ABSTRACT

A method detects similar objects in a collection of such objects by modification of a previous method in such a way that per-object memory requirements are reduced while false detections are avoided approximately as well as in the previous method. The modification includes (i) combining k samples of features into s supersamples, the value of k being reduced from the corresponding value used in the previous method; (ii) recording each supersample to b bits of precision, the value of b being reduced from the corresponding value used in the previous method; and (iii) requiring l matching supersamples in order to conclude that the two objects are sufficiently similar, the value of l being greater than the corresponding value required in the previous method. One application of the invention is in association with a web search engine query service to determine clusters of query results that are near-duplicate documents.